Audio-Visual Based Online Multi-Source Separation

Jonah Ong, Ba Tuong Vo, Sven Nordholm, Ba-Ngu Vo, Diluka Moratuwage, and Changbeom Shim

***Abstract*—Meeting or conference assistance is a popular application that typically requires compact configurations of co-located audio and visual sensors. This paper proposes a novel solution for online separation of an unknown and time-varying number of moving sources using only a single microphone array co-located with a single visual device. The approach exploits the complementary nature of simultaneous audio and visual measurements, accomplished by a model-centric 3-stage process of detection, tracking, and (spatial) filtering, which performs separation in a block-wise or recursive fashion. Fusing the measurements requires solving the multi-modal space-time permutation problem, since the audio and visual measurements reside in different observation spaces, but also are unidentified or unlabeled (with respect to the unknown and time-varying number of sources), and are subject to noise, extraneous measurements and missing measurements. A labeled random finite set tracking filter is applied to resolve the permutation problem and recursively estimate the source identities and trajectories. A time-varying set of generalized side-lobe cancellers is constructed based on the tracking estimates to perform online separation. Evaluations are undertaken with live human speakers.**

***Index Terms*—Audio-visual, source separation, spatial filtering, labeled random finite sets, generalized labeled multi-Bernoulli**

I. Introduction and Related Works

SOURCE separation refers to the estimation of individual source signals from an unknown mixture signal recorded by one or more microphones. A common challenge in source separation is the permutation ambiguity problem [1]. Traditional approaches to blind source separation (BSS) such as independent component analysis (ICA) [2], sparseness-based solutions [3], [4] and non-negative matrix factorization (NMF) [5] have demonstrated strong interference suppression with minimal signal distortion on a mixture of static speech sources. These approaches typically assume a fixed and known number of stationary sources and exploit their individual statistics in order to achieve separation. More recent deep neural network (DNN) based approaches such as uPIT [6], DPCL [7], DANet [8], and TasNet [9] have also shown promising separation performance for pre-trained speaker models. Similarly, these approaches rely on the assumption that the number of speakers and their characteristics are fixed and known during training and testing [10].

Source separation for an unknown and time-varying number of moving speakers is even more challenging since the room impulse response for each source varies in both time and position [11]. As a result, standard BSS techniques which rely on stationarity assumptions may not be directly applicable [12], [13]. In addition, it is not clear if DNN based approaches can be extended to accommodate the unknown and spontaneous appearance and disappearance of active sources. An alternative to BSS and DNN approaches is a model-centric approach based on a 3-step process of detection, tracking, and filtering (DTF), which has the salient feature of being able to accommodate an unknown and time-varying number of moving sources without pre-training [13]–[15]. In our previous work [16], the DTF approach was further demonstrated for online or recursive operation using the latest generation of random finite set (RFS) tracking techniques [17], [18], where separation of multiple speech sources was achieved through initially taking audio measurements from multiple microphone arrays, then tracking the sources in space and time, and finally carrying out beamforming in the direction of the estimated source. While each of the abovementioned approaches has relative advantages and disadvantages in different applications, the common element is that they exclusively rely on audio content to perform separation.

In noisy or loud settings, humans can employ both audio and visual cues to hone in on the speaker of interest, and are thought to incorporate the audio-visual correspondence between lip movements and speech utterances [19]. Motivated by traditional BSS approaches, an unsupervised audio-visual solution is proposed in [20], which employs low-rank matrices to model the background audio-visual information, while sparsity is used to extract sources through correlations between the audio and visual modalities. The DNN-based solution proposed in [21] uses an off-the-shelf face detector in combination with a face recognition model to extract face embeddings and estimate the associations of speech signals to their respective speakers. Subsequent works in [10], [22] incorporate a DNN-module that extracts lip embeddings and facial appearance directly from video streams, exploiting joint audio-visual features in matching lip movements and voice fluctuations to the correct speaker. The work in [23] further analyzes the close connection between facial motion and emitted speech, proposing that the consistency between voice elements and facial appearance can facilitate the isolation of speech from overlapping sounds.

DNN-based solutions for audio-visual source separation have also been specialized to exploit the naturally occurring features in the case of musical sources. Live musical sounds typically emanate from a person playing an instrument with a unique action, and it is possible to exploit the distinctive correspondence between the audio and visual cues of music generation to achieve separation. To date, numerous DNN-based solutions have shown promising audio-visual based separation performance. The work in [24] demonstrates that a mix of different musical instruments playing on video can

The authors are with the Department of Electrical and Computer Engineering, Curtin University, Australia. (e-mail: jonahosx25@gmail.com; ba-tuong.vo@curtin.edu.au; S.Nordholm@curtin.edu.au; ba-ngu.vo@curtin.edu.au; Diluka.Moratuwage@curtin.edu.au; Changbeom.Shim@curtin.edu.au). This work is supported by the Australian Research Council under FT210100506 and LP200301507.

be separated by locating the cluster of pixels corresponding to the sound from a particular instrument. This method exploits the natural synchronization of audio and visual modalities to enable joint audio-visual learning without supervision [24], and was extended to train a self-supervised network for vehicle tracking with stereo sound [25].

When multiple similar instruments are playing, relying solely on audio and visual semantics is typically insufficient. The more recent solution in [26] additionally incorporates temporal motion information from the video to improve source differentiation and hence sound separation. An alternative approach in [27] considers the correspondence between body dynamics and finger movements to create a context-aware network which enables more robust audio-visual separation of both heterogeneous and homogeneous musical sources. Network training can further be improved with a so-called sounding object visual grounding technique [28], which distinguishes between active and silent sources to avoid learning noise from the latter. Noting that simultaneous musical instruments are usually interactive in their timing, the approach in [29] improves on one-time separation solutions by recursively minimizing the residual information in the spectrogram. DNN-based audio-visual solutions have also found applications in robot navigation [30], [31], automatic speech recognition [32]–[34], and person recognition [35]–[38].

The abovementioned approaches to audio-visual based separation are broadly classified as being data-centric, in the sense that they require some form of training to capture the correspondence between the two complementary modes. Data-centric approaches generally rely on large training sets to work desirably [21], [23] which can be computationally intensive during the learning stage. Moreover, the abovementioned data-centric approaches are generally regarded as offline or batch methods, as the output decompositions are produced only after processing the entire input history, as opposed to online methods where the output and input are synchronized up to a fixed delay. In addition, it is not immediately clear if such approaches are amenable for the separation of an unknown and time-varying number of moving sources.

In contrast to data-centric, model-centric DTF approaches to audio-visual based separation are virtually unexplored. The use of co-located audio and visual sensors is intuitively appealing since the two complementary modalities are used to observe the same scene. This approach is also naturally suited to online conferencing or meeting analysis type applications, where both modes are readily available and are likely to be more effective than using audio data alone. One of the main difficulties lies in fusing the two measurement modes since the 3D audio measurements and 2D video measurements reside in different observation spaces even though they observe the same physical space or state space. Furthermore, the audio and visual measurements are subject to noise, spurious or missing measurements, and are unlabeled or unidentified. In addition, active sources can move, while new sources can appear and existing sources can disappear. Collectively, these issues give rise to the *multi-modal space-time permutation problem*, since it is not known which measurements are connected to which sources (if any at all) in both measurement modes and across space and time.

Multi-source separation becomes far more challenging in the popular commercial application of meeting or conference assistance. Such applications require a compact configuration with a small number of co-located audio and visual sensors for spatial efficiency and portability as well as facilitating synchronization and calibration [39]. The ensuing technological question is whether multi-source separation can be achieved with this minimal configuration. Apart from the low observability, the absence of widely spaced sensors reduces the available spatial information, thereby causing more noise in the measurements [40]. A co-located sensor configuration therefore relies on the complementarity of both modalities to yield accurate tracking results and improve source separation. Intuitively, visual observations are used to reduce the uncertainty in 3D localization and assist the audio measurement [41], [42], which facilitates better directionality and suppression in spatial filtering.

This work proposes a novel model-centric DTF based algorithm for online source separation, using only a single microphone array co-located with a single visual device. The proposed approach caters for an unknown and time-varying number of moving sources, without pre-training, by exploiting the complementary nature of simultaneous audio and visual measurements. An RFS framework [17], [18] is adopted to address the fusion of the multi-modal measurements and to facilitate the tracking of multiple moving sources. The RFS approach entails the development of stochastic models which capture the physical relationship between the measurements and the sources, including the abovementioned uncertainties. An RFS tracking filter known as the Multi-Sensor Generalized Labeled Multi-Bernoulli (MS-GLMB) filter [43]–[46] is applied to recursively estimate the number of sources as well as their identities and trajectories, thereby addressing the multi-modal space-time permutation problem. The tracking estimates inform the construction of a time-varying set of spatial filters, known as Generalized Side-lobe Cancellers (GSCs) [47] for achieving source separation. Near-field and far-field evaluations are undertaken with live human speakers.

In summary, our main contribution is a novel audio-visual source separation algorithm, which is the first to demonstrate

- Model-based solution via detection, tracking and filtering,
- Operation in an online fashion or as the data arrives,
- An unknown time-varying number of moving sources,
- Separation without pre-training of the audio signals.

II. Problem Formulation and Solution Overview

A. Signal Model

Consider a scenario where the number of sources is time-varying, and let $N(t)$ denote the number of sources in the scene at discrete time instance t. Each source indexed by $n \in \{1, ..., N(t)\}$ is located at position vector $\alpha_n(t) \in \mathbb{R}^3$ at the time instance t. The signal emitted by source n is denoted by s_n, and all signals are assumed to be mutually uncorrelated. The source signals propagate and are received by a single array of M microphones, where each microphone element indexed

by $m \in \{1, ..., M\}$ is corrupted with non-directional diffuse noise $v^{(m)}$. In this work, we assume source stationarity at each frame k of length T, i.e. $\alpha_n(t) = \alpha_{k,n}$ and $N(t) = N_k$ for $t = (k-1)T, ..., kT$. In this case, the source signal s_n can be represented in blocks of frames:

$$s_n(t) = \sum_{k=1}^{K} s_n(t) w_T\left(t - (k-1)T\right) = \sum_{k=1}^{K} s_{k,n}(t), \quad (1)$$

where w_T is a window function of length T, and k is the index of a time block/frame with length T. Based on the direct path term only, the mixture received by microphone element m is approximated by:

$$y^{(m)}(t) \approx \sum_{k=1}^{K} \sum_{n=1}^{N_k} \frac{s_{k,n}\left(t - \tau(\alpha_{k,n}, u^{(m)})\right)}{4\pi ||\alpha_{k,n} - u^{(m)}||} + v^{(m)}(t), \quad (2)$$

where $||\cdot||$ is the Euclidean distance, $\tau(\alpha_{k,n}, u^{(m)}) \triangleq c_{\mathrm{s}}^{-1} ||\alpha_{k,n} - u^{(m)}||$ is the time delay between source n at position $\alpha_{k,n}$ and microphone m at position $u^{(m)} \in \mathbb{R}^3$, and c_{s} is the speed of sound propagation. The objective is to estimate the individual source signals frame by frame using the mixture signals $y^{(1)}, ..., y^{(M)}$ with no prior knowledge on the number of sources, their positions and identities/labels.

B. Visual Assistance

To estimate the individual source signals, knowledge of the source positions and their labels is crucial, as they are needed to direct a set of time-varying spatial filters to perform source separation. In our previous work [16], this is achieved by tracking multiple sources in 3D space using audio-only data obtained from four microphone arrays that are spaced around the room. The use of multiple microphone arrays is necessary because the audio measurements obtained from a single array alone typically have insufficient observability to allow accurate 3D tracking. An alternative to multiple microphone arrays is to use complementary audio-visual data to observe multiple human speakers in a common physical space. According to recent surveys [48], [49], visual detections or measurements via standard object detectors, e.g. body [50], face [51], and pose [52], have become highly robust and accurate over the years. Thus the use of a single visual device in combination with a single microphone array is likely to facilitate accurate tracking performance. Due to the complementary nature of the audio and visual measurements, which are conditionally independent measurements of the same active sources in a common physical space, it is natural to exploit both modalities simultaneously. To incorporate 2D visual measurements with 3D audio measurements, it is necessary to specify the physical relationship $\mathcal{P}_V^{(c)}$, which maps the 3D source position α to the 2D camera projection $\alpha_V^{(c)}$. Details of this relationship are given in the next section.

C. Overview of the Proposed Method

The processing chain of the proposed method is shown in Fig. 1. Audio and visual measurements of the same (multiple) sources in a common (physical) space are synchronized and segmented into frames indexed by discrete time $k = 1, ..., K$. At each frame, raw microphone signals are fed into an acoustic localization technique to acquire the 3D source position candidates. In parallel, images from multiple cameras are fed into a monocular face detection algorithm to acquire 2D centroid measurements of the same sources present. Measurements acquired from both modalities are subjected to noise (disturbance), they may not reflect a source that is present (false negative), and some may not correspond to any source (false positive). Furthermore, the audio and visual measurements undergo different transformations and hence reside in different observations spaces. Consequently, the audio and visual measurements have an inherent *multi-modal space-time permutation* issue, since the measurements are unlabeled or unidentified with respect to the time-varying and unknown number of sources. The space permutation aspect refers to the fact that in a given frame, it is not known which measurements (if any) correspond to which sources, while the time permutation aspect refers to the fact that across time, it is not known which measurements (if any) correspond to the same source. A labeled RFS approach [43]–[46] can be used to model the stochastic relationship between the multi-modal measurements and source states, and jointly estimate the number of sources, their positions and labels. Based on the tracking estimates, a set of time-varying spatial filters can be constructed based on the direct path signal model to perform source separation. The proposed method can be described in three stages: audio-visual measurement acquisition, multi-source tracking, and source separation.

1) Audio-Visual Measurement Acquisition: In the first stage, audio measurements are obtained by first performing the Short-Time Fourier Transform (STFT) on the raw microphone signals. For each frame, the Steered-Response Power Phase Transform (SRP-PHAT) and a region search algorithm known as Stochastic Region Contraction (SRC) [53], are used to obtain 3D position candidates from the microphone array. In parallel, visual measurements are obtained by passing images into the Dual-Shot Face Detector (DSFD) [51] to acquire visual detections in the form of bounding boxes, and then picking the centroids as 2D position candidates of the human lips.

2) Multi-Modal Multi-Source Tracking: In the second stage, we adopt a labeled RFS framework [43]–[46] to fuse the multi-modal (audio-visual) measurements, and produce estimates of the 3D source positions and labels at each frame, in a statistically consistent manner. In this framework, the relationship between the multi-modal measurements and multi-source states is established by the multi-sensor audio-visual measurement model. The motion, appearance, and disappearance of sources are encapsulated by the multi-source transition model. Specifically, a tracking filter known as the Multi-Sensor Generalized Labeled Multi-Bernoulli (MS-GLMB) filter [46] is employed. The recursive filter propagates a so called filtering density, which provides a stochastic description of the set of labeled source states at the current time frame, given all audio-visual measurements up to the current time frame. An estimator is applied to the filtering density to output the source positions and labels at each frame.

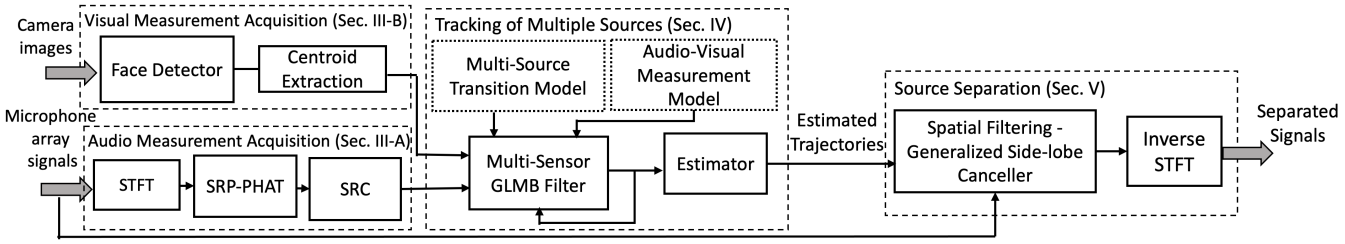

Fig. 1: System Diagram.

3) Source Separation via Spatial Filtering: In the third stage, source separation is achieved via constructing a type of spatial filter known as the Generalized Side-lobe Canceller (GSC) [47]. A set of GSCs is constructed, one for each source present, using the estimated source positions and the labels at each frame. Each GSC is employed to emphasize each source of interest while simultaneously suppressing other interfering sources. Finally, the time-domain separated signals are recovered using the inverse STFT.

III. Audio-Visual Data Pre-Processing

A. Audio Measurement Acquisition

Each raw microphone signal $y^{(m)}$ is segmented into $y_1^{(m)}, ..., y_K^{(m)}$ via:

$$y_k^{(m)}(t) = y^{(m)}(t + (k-1)T) w_T(t)\,, \tag{3}$$

where w_T is a selected window function of length T. The window function is chosen to capture enough signal information while reducing signal discontinuities at the edges, e.g. a Hann window $w_T(t) = 0.5 - 0.5\cos(2\pi t/T),\ \ t = 0, \ldots, T-1$.

We denote the discrete STFT of $y_k^{(m)}$ by $Y_k^{(m)}$. To represent the segmented frequency-domain raw signals from all microphones in a compact form, we stack them into a vector (where λ is the frequency bin index):

$$Y_k(\lambda) = \left[Y_k^{(i)}(\lambda) \right]_{i=1}^{M}. \tag{4}$$

Given Y_k received at the array, the spatial power that emanates from the direction of the source location $\alpha_k \in \mathbb{R}^3$, is computed using SRP-PHAT by [53]:

$$\mathcal{P}_{A,k}(\alpha) = \sum_{a=1}^{M-1} \sum_{b=a+1}^{M} \sum_{\lambda} \frac{Y_k^{(a)}(\lambda) Y_k^{*(b)}(\lambda)}{\left| Y_k^{(a)}(\lambda) Y_k^{*(b)}(\lambda) \right|} \times e^{j\omega_\lambda \left(\tau(\alpha, u^{(b)}) - \tau(\alpha, u^{(a)}) \right)}, \tag{5}$$

where $\omega_\lambda = 2\pi(\lambda - 1)F_s/T$, F_s is the sampling frequency, the PHAT weighting is frequency-dependent, and the exponential term time-aligns the microphone signals based on the propagation delays. Using the computationally efficient SRC algorithm [53], the 3D source position candidates are obtained via peak-picking on SRP-PHAT with a certain threshold (see Fig. 2). We denote the collection of the 3D position candidates as a measurement set:

$$Z_{A,k} = \{ z_{A,k,1}, ..., z_{A,k,|Z_k|} \}, \tag{6}$$

where $|Z_{A,k}|$ denotes the number of elements of $Z_{A,k}$.

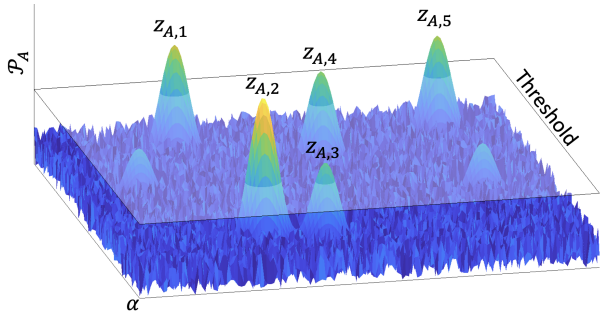

Fig. 2: SRP-PHAT Measurements.

B. Visual Measurement Acquisition

Objects in the 3D world coordinate frame are observed by multiple cameras indexed by $c \in \{1, ..., C\}$, wherein each camera produces object detections as 2D points in the camera image coordinate frame. Each camera is treated as a projective device that converts 3D world points onto the 2D image plane [54]. The perspective projection of a point in the 3D coordinate frame (world) to a point in a 2D coordinate frame (plane) is a nonlinear transformation because it can be interpreted as a many-to-one morphism $\mathbb{R}^3 \to \mathbb{R}^2$ (except for an orthographic projection). Alternatively, this projection can be realized as a linear transformation in the homogeneous coordinates of the projective space $\mathbb{P}$, which is an extension of Euclidean space by adding an extra dimension [54].

Let $\mathcal{P}_V^{(c)}$ be the projective transformation of camera c that takes an arbitrary point α in 3D to a point $\alpha_V^{(c)}$ in 2D (see Fig. 3). Based on the pinhole camera model [54], the transformation $\mathcal{P}_V^{(c)}$ first converts the vector $\alpha = (\alpha_1, \alpha_2, \alpha_3)^T$ into its homogeneous form $\tilde{\alpha} = (\alpha_1, \alpha_2, \alpha_3, 1)^T$ (where the subscript indexes refer to the respective coordinate values), and then performs a linear transformation via the camera matrix $\mathrm{P}_{3\times4}^{(c)}$ to obtain the projected homogeneous point $\tilde{\alpha}_V^{(c)}$ on camera c, i.e.

$$\tilde{\alpha}_V^{(c)} = \mathrm{P}_{3\times4}^{(c)} \tilde{\alpha}. \tag{7}$$

The actual 2D point on the image plane $\alpha_V^{(c)}$ is recovered via dividing the first two coordinate values of $\tilde{\alpha}_V^{(c)} = (\tilde{\alpha}_{V,1}^{(c)}, \tilde{\alpha}_{V,2}^{(c)}, \tilde{\alpha}_{V,3}^{(c)})^T$ by the value of its last coordinate, i.e.

$$\alpha_V^{(c)} = (\tilde{\alpha}_{V,1}^{(c)} / \tilde{\alpha}_{V,3}^{(c)}, \tilde{\alpha}_{V,2}^{(c)} / \tilde{\alpha}_{V,3}^{(c)})^T. \tag{8}$$

The camera matrix $\mathrm{P}_{3\times4}^{(c)}$ of camera c captures the intrinsic parameters (the focal length, skew coefficient and projection

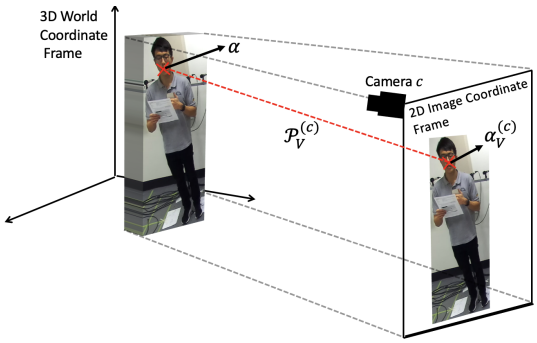

Fig. 3: Projective transformation $\mathcal{P}_V^{(c)}$ of a point α in 3D to a point $\alpha_V^{(c)}$ in 2D for camera c.

center), and the extrinsic parameters (the rotation and translation of the camera), which are obtainable via standard camera calibration techniques [55].

Denote the image obtained from camera c at time frame k by $\mathcal{I}_k^{(c)}$. The image is fed into a Dual-Shot Face-Detector [51] which is represented as detection operator $\mathcal{D}^{(c)}$ and produces a set of 2D visual detections at frame k:

$$Z_{V,k}^{(c)} = \mathcal{D}^{(c)}(\mathcal{I}_k^{(c)}) = \{z_{V,k,1}^{(c)}, ..., z_{V,k,|Z_{V,k}^{(c)}|}^{(c)}\}, \tag{9}$$

where $z_{V,k}^{(c)} = (\alpha_{V,k,1}^{(c)}, \alpha_{V,k,2}^{(c)})^T$ is a point specified in 2D image coordinates, $|Z_{V,k}^{(c)}|$ denotes the number of visual measurements at camera c. Note that the projective transformation $\mathcal{P}_V^{(c)}$ between a 3D point in world coordinates and the observed point in 2D image coordinates establishes the relationship between the 3D source positions and the 2D visual measurements.

C. Audio-Visual Measurements

The multi-modal measurements Z_k at frame k comprise all the constituent measurement sets from the audio and visual sensors, i.e.

$$Z_k = (Z_{A,k}, Z_{V,k}), \tag{10}$$

where $Z_{V,k} \triangleq (Z_{V,k}^{(1)}, ..., Z_{V,k}^{(C)})$. The multi-modal measurements are the basis for estimating the states and labels of the sources. However, the following difficulties arise in the estimation:

- While the audio and visual sensors observe the same scene and same sources, the individual audio measurements $z_{A,k} \in Z_{A,k}$ and individual visual measurements $z_{V,k}^{(c)} \in Z_{V,k}^{(c)}$ are in different observation spaces.
- Due to undergoing different and highly non-linear transformations, individual measurements are noisy, and each measurement set may contain false positives (measurements not generated by any source) and false negatives (missing measurements or missed detections).
- These factors give rise to the inherent *multi-modal space-time permutation problem*, since in space it is not known how the audio measurements from $Z_{A,k}$ and the visual measurements from $Z_{V,k}^{(1)}, ..., Z_{V,k}^{(C)}$ are associated, or generated by which source if any; and in time, it is not known how the individual audio and visual measurements from $Z_{A,k}$ and $Z_{V,k}^{(1)}, ..., Z_{V,k}^{(C)}$ at the current frame are connected to those from $Z_{A,k+1}$ and $Z_{V,k+1}^{(1)}, ..., Z_{V,k+1}^{(C)}$ at the next frame.

In the next section, we show how the multi-modal space-time permutation problem can be solved using a dynamic Bayesian estimation framework. A labeled RFS model [43]–[46] facilitates a statistically consistent specification of the *multi-source transition model* and the *multi-modal measurement model*. The transition model is given by a transition density that captures the appearance, disappearance and motion of the sources over time, and captures the uncertainties due to the time permutation issue. The measurement model is given by a likelihood which is based on the assumption that the audio and visual measurements are conditionally independent given the source states, since the audio and video sensors produce complementary measurements of the same sources in a common scene. Consequently, the audio-visual measurement likelihood is separable and can be written as a product of the audio likelihood and visual likelihood. The audio likelihood function describes the relationship between the SRP-PHAT measurements and the source positions, including the uncertainties due to the space permutation issue. The visual likelihood function describes the relationship between the DSFD measurements and the source positions, based on the pinhole camera model, including the uncertainties due to the space permutation issue. Based on these stochastic transition and measurement models, a Bayesian RFS filter recursively estimates the source trajectories and labels.

IV. Tracking of Multiple Sources

A. Multi-Source Bayes Tracking Filter

The Bayesian RFS framework [17], [18], [56] facilitates the stochastic modeling of the time-varying nature of the number of sources and the individual source positions, as well as the stochastic modeling of the time-varying nature of the number of measurements which are subjected to noise, false measurements (false positives) and missing measurements (false negatives). In tracking terminology, false negatives and false positives are termed missed detections and false detections respectively, while source appearance and disappearance are termed birth and death respectively. The multi-modal space-time permutation problem is referred to as the data association problem and can be addressed using a labeled RFS tracking filter [43]–[46]. A visual illustration of the nature of the multi-modal measurements along with the desired tracking result is shown in Fig. 4.

Each source at frame k has a state denoted by $\boldsymbol{x}_k \triangleq (x_k, \ell_k)$, where $x_k \triangleq (\alpha_k, \dot{\alpha}_k)$ is a vector capturing the 3D position and velocity of the source, and ℓ_k is a unique label from a discrete space $\mathbb{L}_{0:k}$. The inclusion of the velocity component is necessary as an auxiliary variable for the specification of the state transition model. At each frame k, the collection of states for multiple sources is represented as a finite set:

$$\boldsymbol{X}_k = \{\boldsymbol{x}_{k,1}, ..., \boldsymbol{x}_{k,N_k}\}, \tag{11}$$

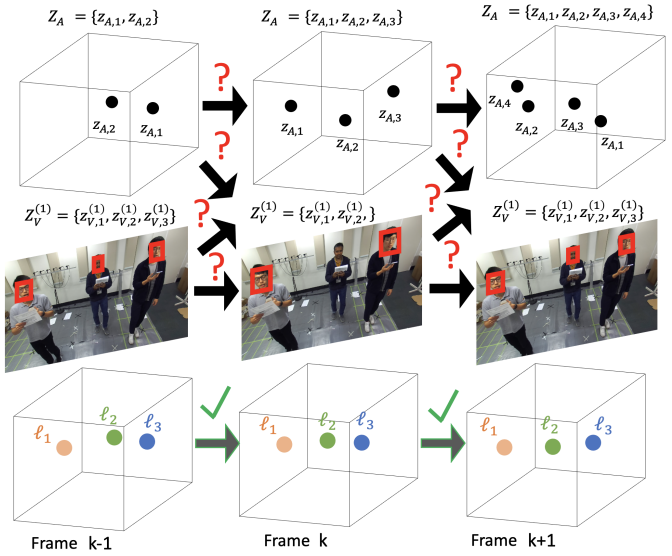

Fig. 4: Three sources existing from frame $k-1$ to $k+1$. The top row shows an illustration of the audio measurements (3D position candidates). The middle row shows an illustration of the visual measurements (2D point detections). Notice the *multi-modal space-time permutation problem,* whereby in space it is not known how audio measurements from $Z_{A,k}$ and visual measurements $Z^{(1)}_{V,k}$ are associated, or generated by which source if any; and in time, it is not known how audio and visual measurements from $Z_{A,k}$ and $Z^{(1)}_{V,k}$ at the current frame are connected to those from $Z_{A,k+1}$ and $Z^{(1)}_{V,k+1}$ at the next frame. The bottom row shows an illustration of the tracking result addressing the multi-modal space-time permutation problem.

herein referred to as a multi-source state, where N_k is the number of sources. A key feature of labeled RFS modeling is the assumption of unique labels in the multi-source state, which treats the trajectory of an individual source as a sequence of states with a common label (see Fig. 4).

In Bayesian RFS filtering, the aim is to estimate frame-by-frame (recursively) the multi-source state $\boldsymbol{X}_k$, given the multi-modal measurements obtained from the beginning of time up to the current time frame k, i.e. $Z_{1:k} \triangleq (Z_1, ..., Z_k)$. The *multi-source Bayes filter* is a recursive mechanism for computing the probability density of $\boldsymbol{X}_k$ given $Z_{1:k}$. In the Bayesian paradigm, such a probability density is referred as the filtering density denoted by $\boldsymbol{\pi}_{k|k}(\boldsymbol{X}_k|Z_{1:k})$, which captures all uncertainty in the multi-source state given $Z_{1:k}$.

The propagation of the filtering density is a recursive procedure consisting of a time-update followed by a data-update. The first step is given by:

$$\boldsymbol{\pi}_{k+1|k}(\boldsymbol{X}_{k+1}|Z_{1:k}) = \int \boldsymbol{f}(\boldsymbol{X}_{k+1}|\boldsymbol{X}_k)\boldsymbol{\pi}_{k|k}(\boldsymbol{X}_k|Z_{1:k})\delta\boldsymbol{X}_k, \quad (12)$$

where the above set integral is derived from Finite Set Statistics (FISST) for dealing with probability densities of RFSs in a mathematically consistent manner [17], [18], and the probability density $\boldsymbol{f}(\boldsymbol{X}_{k+1}|\boldsymbol{X}_k)$ is the *multi-source transition density* or the probability density that multi-source state $\boldsymbol{X}_k$ at frame k transitions to $\boldsymbol{X}_{k+1}$ at the next frame $k+1$. The *multi-source transition density* is derived from a stochastic model that captures all possible source births, deaths and motions, i.e. the previously discussed time permutation aspect. The parameters for the transition model are given in Section IV-B.

The time-updated density (or predicted density) (12) describes the uncertainty in $\boldsymbol{X}_{k+1}$, given all multi-modal measurements $Z_{1:k}$ up to the current time frame, and addresses the time permutation part of the data association problem.

The second step is given by:

$$\boldsymbol{\pi}_{k+1|k+1}(\boldsymbol{X}_{k+1}|Z_{1:k+1}) = \frac{g(Z_{k+1}|\boldsymbol{X}_{k+1})\boldsymbol{\pi}_{k+1|k}(\boldsymbol{X}_{k+1}|Z_{1:k})}{\int g(Z_{k+1}|\boldsymbol{X}_{k+1})\boldsymbol{\pi}_{k+1|k}(\boldsymbol{X}_{k+1}|Z_{1:k})\delta\boldsymbol{X}_{k+1}}, \quad (13)$$

where the probability density $g(Z_{k+1}|\boldsymbol{X}_{k+1})$ is the *multi-modal (audio-visual) measurement likelihood* or the probability density of the multi-modal measurements Z_{k+1} given the multi-source state $\boldsymbol{X}_{k+1}$. The *multi-modal measurement likelihood* is derived from a stochastic model that encapsulates noise, detections, missed detections, false detections and multi-modal association uncertainty, i.e. the previously discussed audio-visual space permutation aspect. The parameters for the multi-modal measurement model are given in Section IV-C. The data-updated density (or new filtering density) (13) contains all information about $\boldsymbol{X}_{k+1}$, conditioned on the multi-modal measurements $Z_{1:k+1}$ up to the new time frame, and addresses the space permutation part of the data association problem.

B. The Standard Multi-Source Transition Model

Given the multi-source state $\boldsymbol{X}_k$, each state $\boldsymbol{x}_k \triangleq (x_k, \ell_k) \in \boldsymbol{X}_k$ either persists and survives with probability P_S and transition to a new state (x_{k+1}, ℓ_{k+1}) that inherits the same label with transition density $f_S(x_{k+1}|x_k, \ell_k)\delta_{\ell_k}[\ell_{k+1}]$, or dies with probability $1-P_S$. The single-source transition density $f_S(x_{k+1}|x_k, \ell_k)$ gives the probability density of source label ℓ_k moving from state x_k to state x_{k+1}. For tracking live human speakers, a popular choice for the transition density is the Langevin model [57]–[59], which takes on the form:

$$f_S(x_{k+1}|x_k, \ell_k) = \mathcal{N}(x_{k+1}; \mathrm{F}x_k, \mathrm{R}\mathrm{R}^T), \quad (14)$$

where $\mathcal{N}(\cdot; \mathrm{F}x_k, \mathrm{R}\mathrm{R}^T)$ is a Gaussian probability density function with mean $\mathrm{F}x_k$ and covariance $\mathrm{R}\mathrm{R}^T$, $\mathrm{F} = \mathrm{F}_{\text{pseudo}} \otimes \mathrm{I}_3$, $\mathrm{R} = \mathrm{R}_{\text{pseudo}} \otimes \mathrm{I}_3$, I_3 an identity matrix of 3 dimensions, $\otimes$ is the Kronecker product,

$$\mathrm{F}_{\text{pseudo}} = \begin{bmatrix} 1 & \phi \\ 0 & e^{-\beta\phi} \end{bmatrix} \quad \mathrm{R}_{\text{pseudo}} = \sigma_{\Xi} \begin{bmatrix} 0 \\ \nu\sqrt{1-e^{-2\beta\phi}} \end{bmatrix}, \quad (15)$$

β is the rate constant that controls the rate at which the velocity decays, ν is the steady-state root-mean-square velocity constant, ϕ is the discretization time step interval, and σ_{Ξ} is a 3D column vector of the component standard deviations of the process noise.

At this next time, a set of new sources denoted by $\boldsymbol{B}_{k+1}$ with labels $\{\ell_{k+1} : (x_{k+1}, \ell_{k+1}) \in \boldsymbol{B}_{k+1}\}$ can appear individually with probability $r_B(\ell_{k+1})$ and distributed according to the birth density $p_B(\cdot, \ell_{k+1})$. A label follows the convention $\ell_k = (\varsigma, \iota) \in \mathbb{L}_k$, where $\varsigma \in \{k\}$ denotes the time frame of birth and $\iota \in \mathbb{N}$ denotes the index of source born at the same time [43]. Consequently, the labels of a multi-source state are

distinct/unique for all frames, and the label space for sources at frame k is constructed recursively by $\mathbb{L}_{0:k}=\mathbb{L}_{0:k-1}\cup\mathbb{L}_k$.

The RFS $\boldsymbol{X}_{k+1}$ is the union of the survivals $\boldsymbol{W}_{k+1}$ and births $\boldsymbol{B}_{k+1}$ which are assumed to be statistically independent. Denote by $\boldsymbol{f}_S(\boldsymbol{W}_{k+1}|\boldsymbol{X}_k)$ and $\boldsymbol{f}_B(\boldsymbol{B}_{k+1})$, the probability densities of the surviving sources $\boldsymbol{W}_{k+1}$ from $\boldsymbol{X}_k$, and the births of new sources $\boldsymbol{B}_{k+1}$ respectively. The *multi-source transition density* is given by [43]:

$$\boldsymbol{f}(\boldsymbol{X}_{k+1}|\boldsymbol{X}_k)=\boldsymbol{f}_S(\boldsymbol{W}_{k+1}|\boldsymbol{X}_k)\boldsymbol{f}_B(\boldsymbol{B}_{k+1})\,. \tag{16}$$

The above product is a stochastic model for addressing the time permutation problem. Under this model, source appearance, disappearance and motion are statistically independent. Importantly, distinct/unique labels are propagated for existing sources that continue to be active. The appearance of new sources is catered for with new distinct labels, while deactivated sources are removed without reusing their labels. The derivation and full expression for (16) is not required for this paper, however readers are referred to the original work [43] for details. The transition density (16) captures all possible source births, deaths and motions in the transition of a multi-source state from one frame to the next, and is parameterized by: the probability of survival P_S, single-source transition density f_S, probability of birth r_B, and the birth density p_B. Specific values for these parameters are provided in the experimental section.

C. The Standard Multi-Sensor Measurement Model

1) Microphone Array Measurements: Given a multi-source state $\boldsymbol{X}_k$, each $\boldsymbol{x}_k=(x_k,\ell_k)\in\boldsymbol{X}_k$ is either detected by the microphone array with probability $P_{A,D}$ and generates a detection $z_{A,k}\in Z_{A,k}$ with a likelihood $g_A(z_{A,k}|x_k,\ell_k)$, or is missed with probability $1-P_{A,D}$. The audio single-source likelihood $g_A(z_{A,k}|x_k,\ell_k)$ gives the probability density of the audio measurement $z_{A,k}$ given the source state (x_k,ℓ_k). For SRP-PHAT measurements, the likelihood has the form:

$$g_A(z_{A,k}|x_k,\ell_k)=\mathcal{N}(z_{A,k};\mathrm{H}x_k,\sigma_A\sigma_A^T), \tag{17}$$

where $\mathrm{H}=[\mathrm{I}_3,0]$, and σ_A is a 3D column vector of the component standard deviations describing the uncertainty in the audio measurement ($\sigma_A\sigma_A^T$ is the 3-by-3 noise covariance matrix).

The detection process also generates false detections, conventionally characterized by an intensity function $\kappa_A(z_{A,k})\triangleq\lambda_A\mathcal{U}_A(z_{A,k})$ on the measurement space [17], [18]. The number of false detections is modeled by a Poisson distribution with mean λ_A, and the false detections themselves are uniformly distributed in the audio measurement space according to $\mathcal{U}_A$. It is standard to assume that the audio detections are statistically independent from the false detections [17], [18].

Let $\mathcal{L}(\boldsymbol{X}_k)$ be a set of all distinct source labels present in $\boldsymbol{X}_k$, i.e. $\mathcal{L}(\boldsymbol{X}_k)\triangleq\{\ell:(x_k,\ell)\in\boldsymbol{X}_k\}$. A single-array association $\theta_{A,k}\in\Theta_{A,k}$ is defined as a mapping from the source labels to the audio measurement indices, i.e. $\theta_{A,k}:\{\ell_k:\ell_k\in\mathcal{L}(\boldsymbol{X}_k)\}\rightarrow\{0:|Z_{A,k}|\}$, such that *no two distinct arguments are mapped to the same positive value* [43]. This property ensures each audio measurement comes from at most one source. For example, $\theta_{A,k}(\ell_k)>0$ corresponds to source ℓ_k generating detection $z_{A,k,\theta_{A,k}(\ell_k)}$, while $\theta_{A,k}(\ell_k)=0$ means a missed detection for source ℓ_k.

The multi-source audio measurement likelihood is given by:

$$g_A(Z_{A,k}|\boldsymbol{X}_k)\propto\sum_{\theta_{A,k}\in\Theta_{A,k}}\prod_{\substack{(x_k,\ell_k)\\\in\boldsymbol{X}_k}}\psi_{A,Z_{A,k}}^{(\theta_{A,k}(\ell_k))}(x_k,\ell_k), \tag{18}$$

where

$$\psi_{A,Z_{A,k}}^{(j)}(x_k,\ell_k)=\begin{cases}\frac{P_{A,D}g_A(z_{A,k,j}|x_k,\ell_k)}{\kappa_A(z_{A,k,j})}, & j>0\\ 1-P_{A,D}, & j=0\end{cases}. \tag{19}$$

The mixture form of the audio measurement likelihood (18) takes in account all possible combinations of missed detections, false detections and the source detections that can occur in the audio measurements.

2) Camera Measurements: Given a multi-source state $\boldsymbol{X}_k$, each $\boldsymbol{x}_k=(x_k,\ell_k)\in\boldsymbol{X}_k$ is either detected by the camera c with probability $P_{V,D}^{(c)}$ and generates a detection $z_{V,k}^{(c)}\in Z_{V,k}^{(c)}$ with a likelihood $g_V^{(c)}(z_{V,k}^{(c)}|x_k,\ell_k)$, or is missed by camera c with probability $1-P_{V,D}^{(c)}$. The visual single-source likelihood $g_V^{(c)}(z_{V,k}^{(c)}|x_k,\ell_k)$ for camera c gives the probability density of the visual measurement $z_{V,k}^{(c)}$ given the source state (x_k,ℓ_k). For 2D camera detections, the likelihood for camera c takes on the form:

$$g_V^{(c)}(z_{V,k}^{(c)}|x_k,\ell_k)=\mathcal{N}(z_{V,k}^{(c)};\mathcal{P}_V^{(c)}(\mathrm{H}x_k),\sigma_V^{(c)}\sigma_V^{(c)T}), \tag{20}$$

where $\mathcal{P}_V^{(c)}$ is the transformation described in Section III-B, and $\sigma_V^{(c)}$ is a 2D column vector of the component standard deviations describing the uncertainty in the visual measurement ($\sigma_V^{(c)}\sigma_V^{(c)T}$ is the 2-by-2 noise covariance matrix).

The detection process also generates false measurements or detections, conventionally characterized by an intensity function $\kappa_V^{(c)}(z_{V,k}^{(c)})\triangleq\lambda_V^{(c)}\mathcal{U}_V(z_{V,k}^{(c)})$ on the measurement space for camera c [17], [18]. The number of false detections is modeled by a Poisson distribution with mean $\lambda_V^{(c)}$, and the false detections themselves are uniformly distributed in the visual measurement space according to $\mathcal{U}_V$. It is standard to assume that the visual detections are statistically independent from the false detections [17], [18].

A single-camera association $\theta_{V,k}^{(c)}\in\Theta_{V,k}^{(c)}$ is defined as a mapping from the source labels to the visual measurement indices, i.e. $\theta_{V,k}^{(c)}:\{\ell_k:\ell_k\in\mathcal{L}(\boldsymbol{X}_k)\}\rightarrow\{0:|Z_{V,k}^{(c)}|\}$, such that *no two distinct arguments are mapped to the same positive value* [43]. This property ensures each visual measurement comes from at most one source. For multiple cameras, a multi-camera association is the vector $\theta_{V,k}\triangleq(\theta_{V,k}^{(1)},...,\theta_{V,k}^{(C)})\in\Theta_{V,k}$ of all single-camera associations having the same aforementioned positive one-to-one property, where $\Theta_{V,k}\triangleq\Theta_{V,k}^{(1)}\times...\times\Theta_{V,k}^{(C)}$ is the space of all possible multi-camera associations [46].

The multi-source visual measurement likelihood is given by:

$$g_V(Z_{V,k}|\boldsymbol{X}_k)\propto\sum_{\theta_{V,k}^{(1)}}...\sum_{\theta_{V,k}^{(C)}}\prod_{\substack{(x_k,\ell_k)\\\in\boldsymbol{X}_k}}\prod_{c=1}^{C}\psi_{V,Z_{V,k}^{(c)}}^{(c,\theta_{V,k}^{(c)}(\ell_k))}(x_k,\ell_k), \tag{21}$$

where $\theta_{V,k}^{(1)} \in \Theta_{V,k}^{(1)}, ..., \theta_{V,k}^{(C)} \in \Theta_{V,k}^{(C)}$, and

$$\psi_{V,Z_{V,k}^{(c)}}^{(c,j)}(x_k,\ell_k)=\begin{cases}\frac{P_{V,D}^{(c)} g_V^{(c)}\left(z_{V,k,j}^{(c)}|x_k,\ell_k\right)}{\kappa_V^{(c)}\left(z_{V,k,j}^{(c)}\right)}, & j>0\\ 1-P_{V,D}^{(c)}, & j=0\end{cases}, \quad (22)$$

The mixture form of the visual measurement likelihood (21) takes in account all possible combinations of missed detections, false detections and the source detections that can occur in the visual measurements.

3) Audio-Visual Measurement Likelihood: While the audio and visual sensors produce different measurements in different observation spaces, they are nonetheless observing the same human speakers in a common physical space. Consequently, the measurement sets from each (audio or visual) sensor can be treated as conditionally independent given the multi-source state, and the multi-modal measurement likelihood at time k can be written as:

$$g(Z_k|\boldsymbol{X}_k) = g_A(Z_{A,k}|\boldsymbol{X}_k) \cdot g_V(Z_{V,k}|\boldsymbol{X}_k). \quad (23)$$

Each constituent likelihood function in (23), i.e. g_A or g_V, contains a nested sum that enumerates all possible associations in that measurement domain, thereby taking into account all possible combinations of missed detections, false detections and the source detections. The product of g_A and g_V in (23) therefore contains all combinations of cross-domain associations, thereby presenting a model for addressing the multi-modal space-time permutation problem.

In summary, the *multi-modal measurement likelihood* describes the statistical connection between the audio measurements $Z_{A,k}$ and the visual measurements $Z_{V,k}$ which are complementary observations of the same state $\boldsymbol{X}_k$. The *multi-modal measurement likelihood* is parameterized by: the audio sensor's probability of detection $P_{A,D}$, single-source likelihood g_A, false detection intensity, κ_A; and the visual sensors' probabilities of detection $P_{V,D}^{(1)}, ..., P_{V,D}^{(C)}$, single-source likelihoods $g_V^{(1)}, ..., g_V^{(C)}$, false detection intensities, $\kappa_V^{(1)}, ..., \kappa_V^{(C)}$.

D. Implementation and State Estimation

The MS-GLMB filter [46] is the analytic solution to the multi-source Bayes recursion (i.e. (12) and (13)) under the standard multi-source transition and multi-sensor measurement models. The filter propagates the time-updated and data-updated filtering densities in a GLMB form:

$$\boldsymbol{\pi}_{k|k}(\boldsymbol{X}_k)=\Delta(\boldsymbol{X}_k)\sum_{\theta_{1:k}\in\Theta_{1:k}}\omega_{k|k}^{(\theta_{1:k})}(\mathcal{L}(\boldsymbol{X}_k))\prod_{\boldsymbol{x}_k\in\boldsymbol{X}_k}p_{k|k}^{(\theta_{1:k})}(\boldsymbol{x}_k), \quad (24)$$

where $\Delta(\cdot)$ is a distinct label indicator, i.e. $\Delta(\boldsymbol{X}_k)=1$ if the cardinality $|\mathcal{L}(\boldsymbol{X}_k)|=|\boldsymbol{X}_k|$, $\theta_{1:k}\in\Theta_{1:k}$ is the history of multi-sensor association mappings up to frame k, i.e. $\theta_{1:k}\triangleq(\theta_1,...,\theta_k)$ where $\theta_k\triangleq(\theta_{A,k},\theta_{V,k})$ and $\Theta_k\triangleq\Theta_{A,k}\times\Theta_{V,k}$. Each $\omega_{k|k}^{(\theta_{1:k})}(\cdot)$ is a non-negative weight such that

$$\sum_{L\subseteq\mathbb{L}_{0:k}}\sum_{\theta_{1:k}\in\Theta_{1:k}}\omega_{k|k}^{(\theta_{1:k})}(L)=1, \quad (25)$$

and can be interpreted as the probability of sources with label set L being active, as well as being associated with the audio and visual measurements given by the association history $\theta_{1:k}$. Each $p_{k|k}^{(\theta_{1:k})}(\cdot,\ell)$ is the probability density of the source state with label ℓ and association history $\theta_{1:k}$.

The MS-GLMB filter offers a polynomial time implementation mechanism, which has a linear complexity in the sum of the total number of measurements across all sensors [46]. At each frame k, the MS-GLMB filter outputs a multi-source state estimate

$$\hat{\boldsymbol{X}}_k=\{(\hat{\alpha}_{k,1},\hat{\ell}_1),...,(\hat{\alpha}_{k,|\hat{\boldsymbol{X}}_k|},\hat{\ell}_{|\hat{\boldsymbol{X}}_k|})\}, \quad (26)$$

via a standard GLMB estimator applied to the GLMB filtering density (24) [46]. The source positions and labels over time constitute the estimated source tracks, thereby resolving the space-time permutation problem that arises from the multi-modal measurements as depicted in Fig. 4.

V. Source Separation

A. Spatial Filtering

The estimate $\hat{\boldsymbol{X}}_k$ acquired at each frame from the tracking filter informs the construction of a set of time-varying beamformers based on a free space direct-path model. We use the GSC [47], which contains two parts: a beamformer that determines the response of the source of interest (SOI), and a blocking mechanism to prevent the SOI from entering the canceler.

To estimate the SOI specified by label $\hat{\ell}_i$, the corresponding beamformer is constructed to achieve two objectives: select the direction of the source specified by the estimated position $\hat{\alpha}_{k,i}$, and suppress other interfering sources specified by $\{(\hat{\alpha}_{k,j},\hat{\ell}_j)\in\hat{\boldsymbol{X}}_k\}_{j=1}^{\hat{N}_k}$ for $i\neq j$, where $\hat{N}_k=|\hat{\boldsymbol{X}}_k|$ is the estimated number of sources. For each time-frequency (TF) point (λ,k), the weight of the beamformer $\hat{W}_{k,\hat{\ell}_i}(\lambda)$ is given by [16]:

$$\hat{W}_{k,\hat{\ell}_i}(\lambda)=\left(\left(\mathrm{D}_{k,\hat{\boldsymbol{X}}_k}(\lambda)\right)^H\right)^{\dagger}r_{\hat{N}_k}(\hat{\ell}_i), \quad (27)$$

where ^{H} is the Hermitian transpose, $\dagger$ denotes the Moore-Penrose pseudo-inverse, $r_{\hat{N}_k}$ is a selection vector whose dimension varies depending on the estimated number of sources $\hat{N}_k$, i.e. $r_{\hat{N}_k}[\hat{\ell}_i]=[\delta_{\hat{\ell}_1}[\hat{\ell}_i],\ldots,\delta_{\hat{\ell}_{\hat{N}_k}}[\hat{\ell}_i]]^T$ such that $\delta_i[j]=1$ if $i=j$ and zero otherwise, and

$$\mathrm{D}_{k,\hat{\boldsymbol{X}}_k}(\lambda)=\begin{bmatrix}e^{j\omega_\lambda\left(\tau(\hat{\alpha}_{k,1},u^{(1)})\right)} & \cdots & e^{j\omega_\lambda\left(\tau(\hat{\alpha}_{k,\hat{N}_k},u^{(1)})\right)}\\ \vdots & \ddots & \vdots\\ e^{j\omega_\lambda\left(\tau(\hat{\alpha}_{k,1},u^{(M)})\right)} & \cdots & e^{j\omega_\lambda\left(\tau(\hat{\alpha}_{k,\hat{N}_k},u^{(M)})\right)}\end{bmatrix}, \quad (28)$$

is a matrix with columns representing the steering vectors for each estimated source. The number of columns depends on the estimated number of sources $\hat{N}_k$. Note that if $\hat{N}_k=1$, (27) reduces to the classical delay-and-sum beamformer.

The blocking matrix is defined to be the orthogonal complement to $\left(\hat{W}_{k,\hat{\ell}_i}(\lambda)\right)^H$ [16], [47]:

$$\mathrm{B}_{k,\hat{\ell}_i}(\lambda)=\mathrm{I}-\hat{W}_{k,\hat{\ell}_i}(\lambda)\left[\left(\hat{W}_{k,\hat{\ell}_i}(\lambda)\right)^H\hat{W}_{k,\hat{\ell}_i}(\lambda)\right]^{-1}\left(\hat{W}_{k,\hat{\ell}_i}(\lambda)\right)^H, \quad (29)$$

where I is an identity matrix. Subsequently, the GSC weight vector is defined by:

$$G_{k,\hat{\ell}_i}(\lambda) = \hat{W}_{k,\hat{\ell}_i}(\lambda) - \mathrm{B}_{k,\hat{\ell}_i}(\lambda)V_k(\lambda), \tag{30}$$

where

$$V_{k,opt}(\lambda) = \arg\min_{V} \sum_{\eta=1}^{k} \gamma^{k-\eta} \left| \left(\hat{W}_{\eta,\hat{\ell}_i}(\lambda) - \mathrm{B}_{\eta,\hat{\ell}_i}(\lambda)V \right)^H Y_\eta(\lambda) \right|^2, \tag{31}$$

$\gamma \in [0,1]$ is a positive constant. Eq. (31) can be solved recursively using Recursive Least Squares (RLS) [60].

The output of the GSC beamformer for the estimated source label $\hat{\ell}_i$ at each TF point (λ, k) is given by:

$$S_{k,\hat{\ell}_i}(\lambda) = \left(G_{k,\hat{\ell}_i}(\lambda) \right)^H Y_k(\lambda). \tag{32}$$

Finally, the estimated time-domain signal $\hat{s}_{\hat{\ell}_i}$ of source label $\hat{\ell}_i$ is given by the inverse STFT.

VI. Experiments

In this section, we present the evaluations for the proposed audio-visual based separation method for live human speakers in an acoustic room. The algorithm is tested in scenarios where human speakers are talking and walking at the same time. We initially consider a detailed analysis of the proposed algorithm in near-field vs far-field. In Scenario 1A, the human speakers are situated closer to the audio-visual sensors, while in Scenario 2, human speakers are situated farther away from the audio-visual sensors. In addition, we present an ablation study for each scenario whereby the measurements, tracking and separation are performed using the audio data only. This is undertaken to demonstrate the improvement in performance due to the combination of audio and visual data. The experimental setup is summarized in Section VI-A, and the parameters used for the proposed algorithm are explained in Section VI-B. The evaluation of the accuracy of the SRP-PHAT measurements is given in Section VI-C, followed by the tracking performance of the MS-GLMB filter in Section VI-D, and the separation performance in Section VI-E. Subsequently in Section VI-F, we consider two additional near-field experiments. Scenario 1B has up to three moving sources appearing at different times, and Scenario 1C has at most one source but with two distinct modes of background interference.

A. Experimental Setup

The experiment is conducted in a $7.67\mathrm{m} \times 3.41\mathrm{m} \times 2.7\mathrm{m}$ enclosed room with reverberation measured at $T_{60} \approx 0.25\mathrm{s}$, using a single linear array of 6 microphones, which are calibrated to the same gain/sensitivity. These microphones are connected into 3 *RME-OctaMic 8-channel* pre-amps. Each pre-amp is daisy-chained via MADI cables into the computer. For the visual sensor, a ZED 2 stereo camera from *StereoLabs* is used to record at 1080p. The linear microphone array and ZED 2 stereo camera are co-located and placed close to the wall of the room as shown in Fig. 5.

Fig. 5: Audio-Visual Sensor Setup.

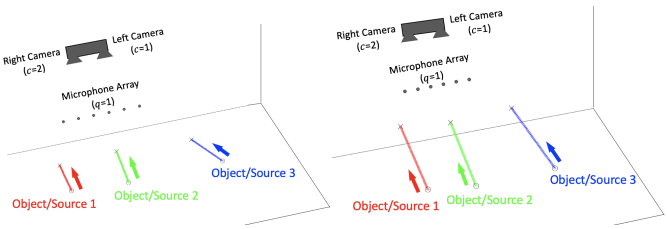

Fig. 6: Scenario 1A (left) and Scenario 2 (right).

To demonstrate the multi-source tracking and source separation performance of the proposed method, Scenario 1A considers three people talking and walking towards the sensors as shown in Fig. 6 (left). The participants stop talking and turn their faces away from the cameras at different times to simulate an exit. A more challenging Scenario 2 employs a similar setup but with the speakers further away from the sensors as shown in Fig. 6 (right). To acquire the original speech signals for evaluation, the participants self-recorded their speech while performing the experiments.

B. Algorithm Parameters

TABLE I: Parameters for microphone array measurements

F_s	16kHz
High-pass filtering	1kHz
Window function	Hann
T	2048
Detector	SRP-PHAT [53]

TABLE II: Parameters for visual device measurements

c	1 (left camera) and 2 (right camera)			
FPS	8			
$\mathrm{P}^{(1)}_{3\times4}$	-1021.7	-827.2	-575.8	7071.5
	31.8	142.0	-1184.0	2012.8
	0.04	-0.83	-0.56	3.81
$\mathrm{P}^{(2)}_{3\times4}$	-1021.9	-822.9	-579.0	6940.6
	28.3	131.3	-1192.6	2030.0
	0.03	-0.82	-0.57	3.81
Detector	Dual-Shot Face Detector [51]			

TABLE III: Parameters for MS-GLMB transition

Multi-source transition density	
β	10s^{-1}
ν	1ms^{-1}
ϕ	0.128s
σ_{Ξ}	$[1.2, 1.2, 0.2]^T\text{ms}^{-1}$
P_S	0.999
$\{r_B(\ell_i)\}_{i=1}^3$	$r_B(\ell_i)=0.005$ for all i
$\{p_B(\cdot,\ell_i) \triangleq \mathcal{N}(\cdot;\mu_B^{(i)}, P_B^{(i)})\}_{i=1}^3$	$\mu_B^{(1)} = [2.0\ 0.7\ 1.7\ 0\ 0\ 0]^T$, $\mu_B^{(2)} = [3.0\ 0.5\ 1.7\ 0\ 0\ 0]^T$, $\mu_B^{(3)} = [4.0\ 0.6\ 1.7\ 0\ 0\ 0]^T$, $P_B^{(i)} = 0.2^2\text{I}_9$ for all i

TABLE IV: Parameters for MS-GLMB likelihood

Audio likelihood	
σ_A	$[0.1, 0.1, 0.1]^T\text{m}$
$P_{A,D}$	0.6
κ_A	$10\mathcal{U}_A$
Visual likelihood	
$\sigma_V^{(c)}$	$[20, 20]^T$ for $c = 1, 2$
$P_{V,D}^{(c)}$	0.99 for $c = 1, 2$
$\kappa_V^{(c)}$	$1\mathcal{U}_V$ for $c = 1, 2$

TABLE V: Parameters for source separation via spatial filtering

Beamformer	Generalized Side-lobe Canceller
Solver	Recursive Least Squares
Window function	Hann
T	2048
Overlap	50%

C. Evaluation of SRP-PHAT Measurements

The audio measurements generated from the single microphone array via SRP-PHAT are in the form of 3D position candidates for active sources. The measurements are not only noisy, but are also subjected to false measurements and missing measurements. To evaluate the accuracy of the audio measurements at each frame, the Optimal Sub-Pattern Assignment (OSPA) metric [61] is applied to quantify the error between the set of audio measurements and the set of true source positions. The OSPA metric typically uses a standard Euclidean distance as a base distance, and a cut-off value beyond which a localization error is deemed to be cardinality error. Consequently, the OSPA metric captures both localization and cardinality errors between the set of measurements and set of truths. The numerical value of the OSPA metric lies between zero and the chosen cut-off, which can be interpreted as a per-point error with units of meters. Further details on the OSPA metric can be found in [61].

The OSPA metric with a cut-off at 1m is shown versus time in Fig. 7 for Scenarios 1A and 2. It can be seen that the error values are consistently high in both scenarios and occasionally saturate at the cut-off value. The time averaged OSPA errors are shown in Table VI, along with the localization and cardinality components. The high average value indicates that the audio-based measurements alone are inaccurate. Furthermore, the large localization component indicates significant positional errors, and the relatively high proportion of the cardinality component indicates significant false and missing measurements. The overall higher errors in Scenario 2 compared to Scenario 1A are due to the sources being farther away from the array. Consequently, the OSPA results in both scenarios suggest that the audio measurements alone are insufficient for accurate tracking of the sources, due to the lack of observability with only a single microphone array.

TABLE VI: Average OSPA distance on the obtained SRP-PHAT measurements.

Scenario	Average OSPA Components (m)		
	Localization	Cardinality	OSPA
1A	0.253	0.561	0.814
2	0.291	0.595	0.886

D. Evaluation of Multi-Source Tracking Filter

The multi-modal audio and visual measurements are modeled in the RFS framework and processed into trajectory estimates with the MS-GLMB filter. The output of the MS-GLMB tracking filter is a set of unique source labels and corresponding position estimates over time which together constitute a set of tracks or trajectories. Due to the imperfect nature of the multi-modal measurements, it is possible that the estimated trajectories will be noisy, in addition to potentially having incorrect labels and/or misaligned starting and finishing times, and extraneous or missing trajectories. To evaluate the accuracy of the audio-visual source tracking, the $\text{OSPA}^{(2)}$ metric [16], [62] can be used, which quantifies the error between the two sets of estimated and true source trajectories. The $\text{OSPA}^{(2)}$ metric uses a time averaged OSPA distance as a base distance between two individual tracks, and has a separate cut-off value beyond which a tracking error is deemed to be a labeling error. Consequently, the $\text{OSPA}^{(2)}$ metric captures both tracking and labeling errors, and the numerical value is interpreted as time-averaged per-track error with units of meters. The metric is typically calculated over a moving window and plotted versus time. Further details on the $\text{OSPA}^{(2)}$ metric can be found in [62].

For this evaluation, a cut-off of 1m is used, with a 10-scan moving window. The $\text{OSPA}^{(2)}$ evaluation for combined audio-visual tracking is shown in Fig. 8 for Scenarios 1A and 2, which for comparison also shows the $\text{OSPA}^{(2)}$ evaluation for audio-only tracking with the single microphone array. It can be seen that in Scenario 1A, combined audio-visual tracking is consistently accurate with low errors below 0.1m. Similarly for Scenario 2, the combination of audio and visual measurements produces consistently accurate tracking estimates with low errors below 0.2m, although the average errors are higher than

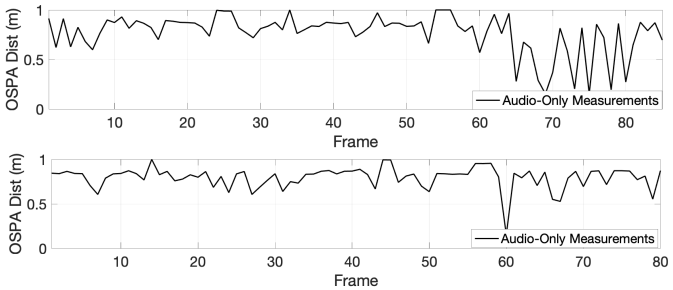

Fig. 7: Scenario 1A (top) and Scenario 2 (bottom): OSPA distance on the SRP-PHAT measurements (lower is better).

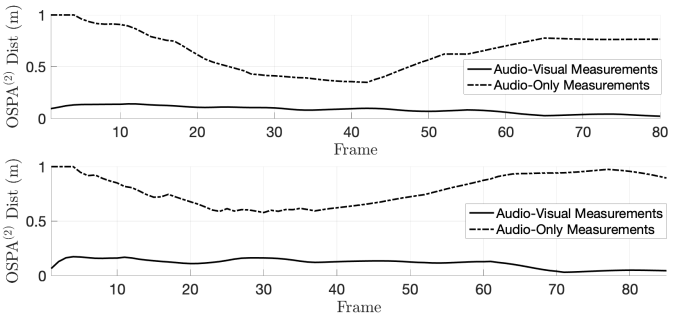

Fig. 8: Scenario 1A (top) and Scenario 2 (bottom): $\mathrm{OSPA}^{(2)}$ distance between estimated and true source trajectories (lower is better).

in Scenario 1A, due to increased distance of the sources from the sensors. Furthermore, the tracking results with only audio measurements from a single microphone array are consistently poor with very high errors in both scenarios. The cause of the relatively high errors for tracking with only audio measurements are not only due to the high positional errors, but also due to label switching errors, and some incidence of extraneous and missing source trajectories. These observations suggest that the multi-modal combination of audio and video measurements enables accurate multi-source tracking, and further highlight the limitations on the observability of the source trajectories with only a single microphone array.

E. Evaluation of Source Separation

The set of position and identity estimates from the MS-GLMB tracking filter are used to perform spatial filtering or source separation via a set of GSCs. As the sources are moving within the room, the delays of each source signal, with respect to the microphone array, are changing over time. Therefore, perceptual measures such as PESQ [63], STOI [64] and PEASS [65], that rely on delay-compensation, are not directly applicable for performance evaluations. While it may be possible to apply these measures on time blocks during which sources are almost stationary, there may be insufficient signal information within each short block to allow a meaningful evaluation [16].

TABLE VII: Scales of SIG, BAK and OVRL in the Subjective Listening Test.

SIG	
Rating	Description
5	Very natural, no degradation
4	Fairly natural, little degradation
3	Somewhat natural, somewhat degraded
2	Fairly unnatural, fairly degraded
1	Very unnatural, very degraded
BAK	
Rating	Description
5	Not noticeable
4	Somewhat noticeable
3	Noticeable but not intrusive
2	Fairly conspicuous, somewhat intrusive
1	Very conspicuous, very intrusive
OVRL	
Rating	Description
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

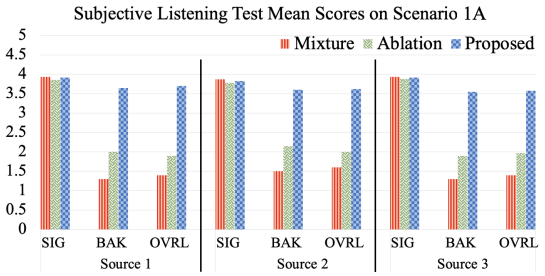

Fig. 9: Mean scores for SIG, BAK, and OVRL for the estimated source signals and original mixture signals evaluated on Scenario 1A.

TABLE VIII: One-way ANOVA test between the estimated source signals and original mixture signals on Scenario 1A.

Source		*p*-value		
		SIG ↑	BAK ↓	OVRL ↓
1	Proposed	0.871*	0.0052	0.0058
	Ablation	0.831*	0.0641*	0.0931*
2	Proposed	0.913*	0.0069	0.0072
	Ablation	0.893*	0.0591*	0.1213*
3	Proposed	0.844*	0.0044	0.0051
	Ablation	0.884*	0.0626*	0.0824*

The asterisk (*) denotes values that are above the selected significance level, i.e. 0.05. (↑ means higher is better while ↓ means lower is better.)

Instead, we administer subjective listening tests based on the ITU-T P.835 methodology which evaluates the extent of signal distortion and the overall quality of noise suppression [66]. In the test, each participant is instructed to listen to the clean speech signal (upper anchor reference), the separated speech signal (to be evaluated) and the mixture signal (lower anchor reference), and then rate them on: The speech signal alone using a five-point scale of signal distortion (SIG); The background interfering noise alone using a five-point scale of background intrusiveness (BAK); The overall quality using a five-point scale of mean opinion score (OVRL). The scales for SIG, BAK and OVRL are described in Table VII.

The evaluation considers the separation performance based on a single microphone array combined with visual tracking assistance from a single camera device (proposed method), and for comparison considers the separation performance using audio-only data without visual tracking assistance (ablation study). In the evaluation, 20 people (12 males, 8 females) of ages from 20 to 30 are recruited to participate in the listening test. A statistical analysis of variance (ANOVA) test at a 0.05 significance level is used to determine if there is a statistically significant difference between the quality of the separated speech signal and the mixture. All video/audio files for both scenarios are available via GitHub: https://github.com/researchwork888/AVseparation.

1) Scenario 1A : Examination of the audio-visual outputs suggests that there is some degree of interference suppression, though the overall performance is naturally constrained by the use of a single microphone array. The mean scores of all 3 criteria, i.e. SIG, BAK and OVRL, are shown in Fig. 9. Some difference is observed in the BAK and OVRL mean scores for all 3 estimated source signals (blue bars) and the mixture signals (orange bars), while the SIG mean scores are relatively similar across the board, which confirms the observed suppression with minimal distortion.

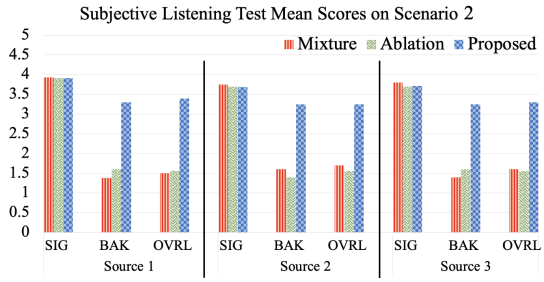

Fig. 10: Mean scores for SIG, BAK, and OVRL for the estimated source signals and original mixture signals evaluated on Scenario 2.

TABLE IX: One-way ANOVA test between the estimated source signals and original mixture signals on Scenario 2.

Source		*p*-value		
		SIG ↑	BAK ↓	OVRL ↓
1	Proposed	0.811*	0.0077	0.0081
	Ablation	0.781*	0.2542*	0.3415*
2	Proposed	0.753*	0.0091	0.0089
	Ablation	0.803*	0.3218*	0.4035*
3	Proposed	0.714*	0.0072	0.0074
	Ablation	0.694*	0.2966*	0.3211*

The asterisk (*) denotes values that are above the selected significance level, i.e. 0.05. (↑ means higher is better while ↓ means lower is better.)

The corresponding *p*-values for the ANOVA test are given in Table VIII. The BAK and OVRL *p*-values for all three sources are below the 0.05 significance value, which suggests a statistically significant difference between the separated and mixture signals in terms of background interference level and overall speech quality. The SIG *p*-values are well above the 0.05 significance level, which suggests that there is no statistically significant difference in terms of signal distortion between the estimated and the mixture signals.

The BAK and OVRL mean scores for the audio-only ablation method (green bars) are much lower than for the proposed audio-visual method, while the SIG mean scores are on par across the board. Furthermore, the BAK and OVRL *p*-values for the ablation are above 0.05 for all sources, which suggests that the audio-only approach produces poor separation performance. In particular, the separated signals produced by the audio-only approach not only have poor interference suppression and overall quality, but are truncated at the start and end of the signals due to late tracking initiation and termination.

Consequently, a co-located audio-visual configuration is capable of performing separation, but is naturally constrained by the limited spatial coverage of the single microphone array. Nonetheless, the use of visual assistance to complement the audio data is still significantly better than an audio-only approach, which is due to vastly improved tracking performance as observed in the previous subsection.

2) Scenario 2: The mean scores of all 3 criteria, i.e. SIG, BAK and OVRL, are shown in Fig. 10, and the results for ANOVA test are given in Table IX. A similar trend is observed to Scenario 1A, although now with lower SIG and BAK scores in Scenario 2. As expected, the proposed audio-visual based approach still achieves a small degree of separation but clearly deteriorates as the sources are placed farther away from the microphones.

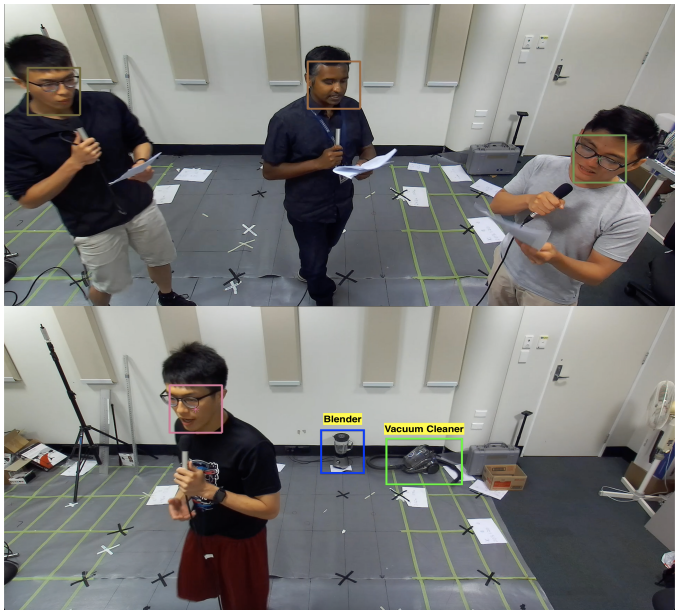

Fig. 11: Screenshots of Scenario 1B (top) and Scenario 1C (bottom).

The results for the audio-only ablation indicate more pronounced failures. The BAK and OVRL means scores for all estimated source signals are low and almost match the scores of the mixture signals. The results of the ANOVA tests also confirm poor separation performance. These failures in the audio-only ablation are expected since the effectiveness of the GSC beamformer is highly dependent on the accuracy of the tracking estimates, which in this case have large localization errors, in addition to extraneous and missing tracks, as well as late initiations and terminations.

In short, while the proposed audio-visual tracking maintains accuracy when sources are farther away, the separation performance degrades with increasing distance between the sources and the single microphone array. However, compared to using audio-only where the separation fails due to erroneous tracking information, the audio-visual approach still maintains consistency in the output.

F. Additional Near-field Experiments

In the previous subsections, it was observed that near-field performance (Scenario 1A) was markedly better than far-field performance (Scenario 2), in all aspects of measurements, tracking, and separation. It was also observed via the ablation studies that audio-visual based separation is much more effective than audio-only separation. We now further explore the audio-visual near-field case with two additional scenarios as described below.

In Scenario 1B, three distinct sources enter the scene at different times, and all are moving while they are speaking. In Scenario 1C, the source enters mid-scenario but its audio is obscured by background noise from a blender and a vacuum cleaner in the room. In both cases the algorithm has no knowledge of the number of sources or the times of their entry. The objective is to separate the mixture of an unknown and time varying number of moving sources.

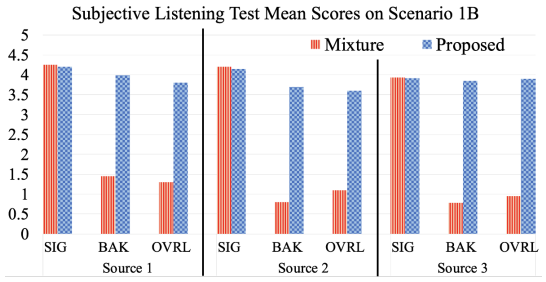

Fig. 12: Mean scores for SIG, BAK, and OVRL for the estimated source signals and original mixture signals evaluated on Scenario 1B.

TABLE X: One-way ANOVA test between the estimated source signals and original mixture signals on Scenario 1B.

Source		*p*-value		
		SIG ↑	BAK ↓	OVRL ↓
1	Proposed	0.891*	0.0057	0.0061
2	Proposed	0.853*	0.0041	0.0044
3	Proposed	0.824*	0.0039	0.0051

The asterisk (*) denotes values that are above the selected significance level, i.e. 0.05. (↑ means higher is better while ↓ means lower is better.)

The screenshots in Fig. 11 illustrate the setup of the two additional scenarios. Due to space constraints we omit the evaluation of the measurements and tracking, as well as the ablation study with audio-only measurements. We only present the evaluation of the separation in a similar manner to Section VI-E. All video/audio files for the additional scenarios are available via GitHub: https://github.com/researchwork888/AVseparation.

1) Scenario 1B (Time-varying Number of Speakers): The mean scores of all 3 criteria, i.e. SIG, BAK and OVRL, are shown in Fig. 12, and the results for ANOVA test are given in Table X. The mean scores and *p*-values of the OVRL and BAK criteria suggest that all three estimated sources achieve good overall speech quality with moderate interference suppression, and similarly the mean scores and *p*-values of the SIG component indicate there is minimal signal degradation or distortion. Additionally, the spectrograms for each of estimated signals are presented in Fig. 13. In this scenario, Source 2 enters the scene a few seconds after Source 1, and Source 3 first appears a few seconds after Source 2. Examination of the spectrograms confirms that the proposed method is able to detect and track all three sources from the point they each enter the scene. As a result, the individual signals for each of the three sources is reconstructed correctly. It is also important to point out that there are no identity switches in the estimation of the trajectories of the sources, which is necessary for the correct reconstruction of the three uninterrupted waveforms. Overall, the results of this scenario demonstrate that the proposed method can handle an unknown and time-varying number of moving sources.

2) Scenario 1C (Loud Background Noise): The mean scores of all 3 criteria, i.e. SIG, BAK and OVRL, are shown in Fig. 14, and the results for ANOVA test are given in Table XI. Additionally, the spectrograms of the obtained signals are presented in Fig. 15. The results indicate that the proposed method is able to detect and track Source 1 quite accurately, and as a consequence, is able to achieve moderate noise suppression with close to no signal distortion. The onset of the source at the two second mark is also correctly initiated with negligible delay, even in the presence of background noise. This is largely due to the exploitation of the complementary audio and visual modes. The results indicate that the proposed method is able to identify the presence, and enhance the speech signal of the moving speaker, with both a blender and vacuum cleaner running simultaneously in the background.

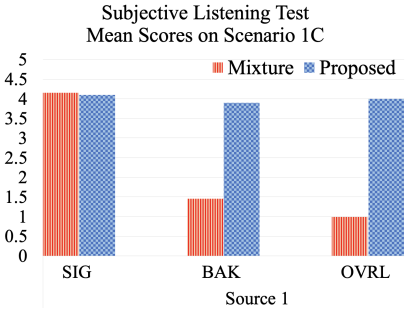

Fig. 14: Mean scores for SIG, BAK, and OVRL for the estimated source signals and original mixture signals evaluated on Scenario 1C.

TABLE XI: One-way ANOVA test between the estimated source signals and original mixture signals on Scenario 1C.

Source		*p*-value		
		SIG ↑	BAK ↓	OVRL ↓
1	Proposed	0.841*	0.0097	0.0088

The asterisk (*) denotes values that are above the selected significance level, i.e. 0.05. (↑ means higher is better while ↓ means lower is better.)

VII. Conclusion

This paper proposes a solution for online separation of an unknown and time-varying number of moving sources, based on a model-centric approach involving sequential stages of detection, tracking, and spatial filtering. The solution exploits simultaneous audio and video measurements, taken from a single microphone array co-located with a single visual device, to produce complementary measurements of an active scene. A labeled random finite set model describes the underlying statistical relationship between the audio-visual measurements and the multi-source states, including the inherent multi-modal space-time permutation uncertainty. A Multi-Sensor GLMB filter is applied to resolve the permutation problem and recursively estimate the source trajectories and labels. A corresponding time-varying set of generalized side-lobe cancellers then performs online source separation.

The proposed solution is evaluated in a real experimental setting with up to 3 live and moving human speakers. An ablation study on audio-only data without the visual mode confirms audibly poor performance due to limited observability with a single microphone array. With the addition of a co-located visual sensor, in near-field experiments, we demonstrate that multi-source separation is possible, despite the limited spatial coverage of the single microphone array. For far-field experiments, the performance is considerably reduced, but still maintains consistency in the output. In both near-field and far-field experiments, the audio-visual approach demonstrably outperforms the audio-only approach. The proposed combination of audio-visual modes is easily extended to the case

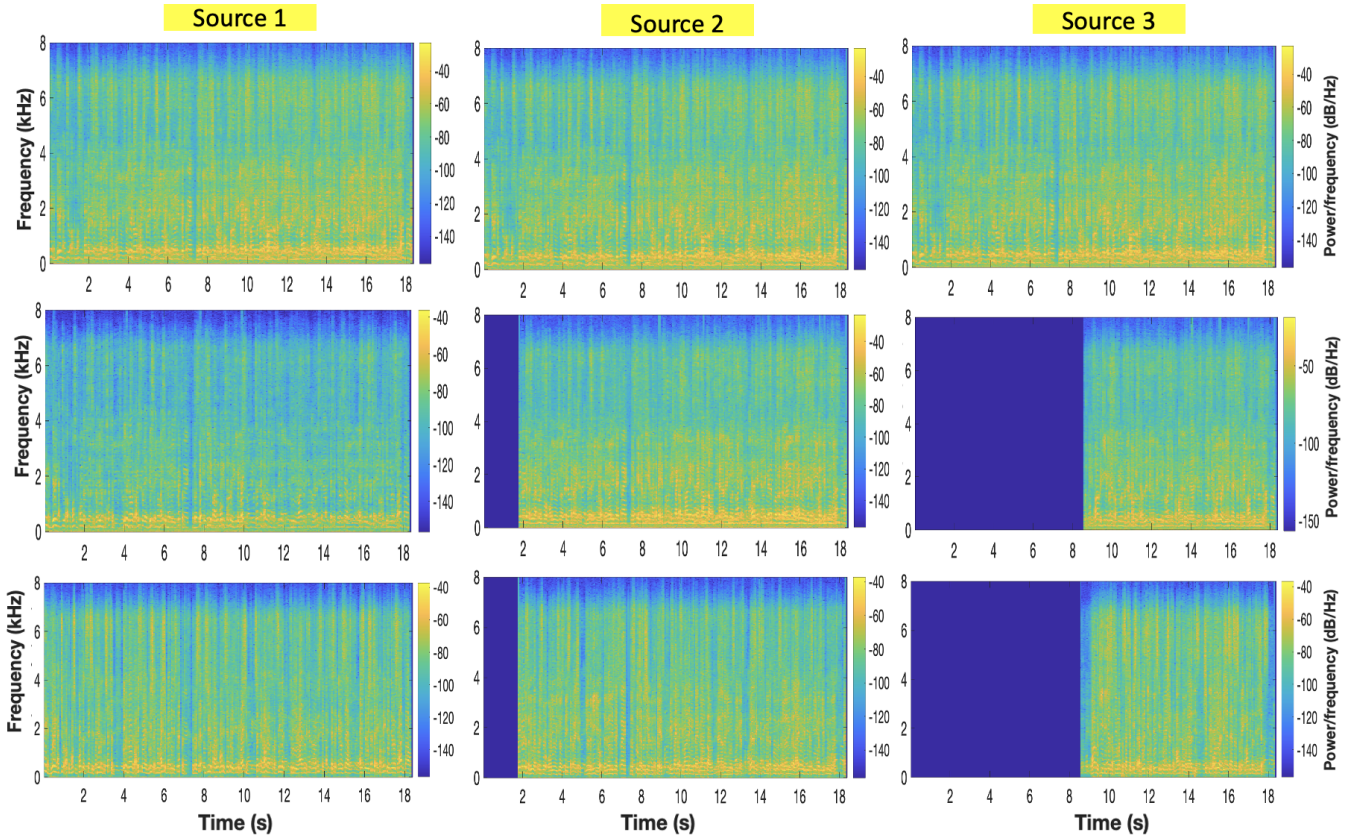

Fig. 13: Spectrograms for signals from Scenario 1B. Top row: mixtures; middle row: estimated signals; bottom row: ground-truth signals.

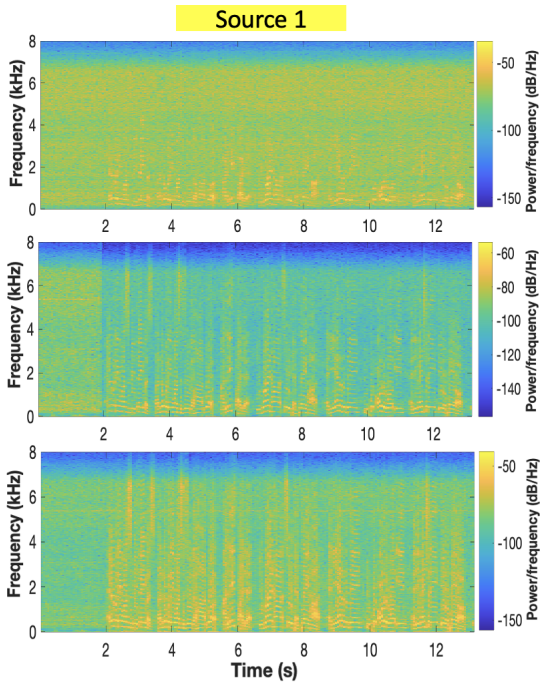

Fig. 15: Spectrograms for signals from Scenario 1C. Top row: mixtures; middle row: estimated signals; bottom row: ground-truth signals.

of multiple visual devices with multiple microphone arrays, which should significantly improve separation performance.

REFERENCES

[1] X. Yu, D. Hu, and J. Xu, *Blind source separation: theory and applications*. John Wiley & Sons, 2013.

[2] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[4] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and-norm minimization," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, p. 024717, 2006.

[5] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.

[6] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[8] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[9] Y. Luo and N. Mesgarani, "TasNet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[10] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 667–673.

[11] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1620–1643, 2020.

[12] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1408–1423, 2016.

[13] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel nmf and acoustic tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 281–295, 2017.

[14] M. Taseska and E. A. Habets, "Blind source separation of moving sources using sparsity-based source detection and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 657–670, 2017.

[15] K. Weisberg, B. Laufer-Goldshtein, and S. Gannot, "Simultaneous tracking and separation of multiple sources using factor graph model," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2848–2864, 2020.

[16] J. Ong, B. T. Vo, and S. E. Nordholm, "Blind separation for multiple moving sources with labeled Random Finite Sets," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[17] R. P. Mahler, *Statistical Multisource-Multitarget Information Fusion*. Artech House, Inc., 2007.

[18] ——, *Advances in Statistical Multisource-Multitarget Information Fusion*. Artech House, 2014.

[19] R. Lu, Z. Duan, and C. Zhang, "Listen and look: Audio–visual matching assisted speech source separation," *IEEE Signal Processing Letters*, vol. 25, no. 9, pp. 1315–1319, 2018.

[20] J. Pu, Y. Panagakis, S. Petridis, and M. Pantic, "Audio-visual object localization and separation using low-rank and sparsity," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2901–2905.

[21] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv preprint arXiv:1804.03619*, 2018.

[22] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

[23] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 495–15 505.

[24] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 570–586.

[25] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7053–7062.

[26] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, "The sound of motions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1735–1744.

[27] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba, "Music gesture for visual sound separation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 478–10 487.

[28] Y. Tian, D. Hu, and C. Xu, "Cyclic co-learning of sounding object visual grounding and sound separation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2745–2754.

[29] M. Shuo, Y. Ji, X. Xu, and X. Zhu, "Vision-guided music source separation via a fine-grained cycle-separation network," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4202–4210.

[30] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, "Look, listen, and act: Towards audio-visual embodied navigation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9701–9707.

[31] J.-T. Lee, M. Jain, H. Park, and S. Yun, "Cross-attentional audio-visual fusion for weakly-supervised action localization," in *International Conference on Learning Representations*, 2020.

[32] G. Sterpu, C. Saam, and N. Harte, "Attention-based audio-visual fusion for robust automatic speech recognition," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 111–115.

[33] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 6548–6552.

[34] F. Tao and C. Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1290–1302, 2018.

[35] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3D convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22 081–22 091, 2017.

[36] G. Sell, K. Duh, D. Snyder, D. Etter, and D. Garcia-Romero, "Audio-visual person recognition in multimedia data from the IARPA janus program," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3031–3035.

[37] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8427–8436.

[38] S. Hörmann, A. Moiz, M. Knoche, and G. Rigoll, "Attention fusion for audio-visual person verification using multi-scale features," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 281–285.

[39] X. Qian, A. Xompero, A. Cavallaro, A. Brutti, O. Lanz, and M. Omologo, "3D mouth tracking from a compact microphone array co-located with a camera," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3071–3075.

[40] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Multi-speaker tracking from an audio-visual sensing device," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2576–2588, 2019.

[41] H. Liu, Y. Sun, Y. Li, and B. Yang, "3D audio-visual speaker tracking with a novel particle filter," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 7343–7348.

[42] X. Qian, A. Brutti, M. Omologo, and A. Cavallaro, "3D audio-visual speaker tracking with an adaptive particle filter," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2896–2900.

[43] B.-T. Vo and B.-N. Vo, "Labeled random finite sets and multi-object conjugate priors," *IEEE Trans. Signal Process.*, vol. 61, no. 13, pp. 3460–3475, 2013.

[44] B.-N. Vo, B.-T. Vo, and D. Phung, "Labeled random finite sets and the Bayes multi-target tracking filter," *IEEE Transactions on Signal Processing*, vol. 62, no. 24, pp. 6554–6567, 2014.

[45] B.-N. Vo, B.-T. Vo, and H. G. Hoang, "An efficient implementation of the generalized labeled multi-Bernoulli filter," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1975–1987, 2017.

[46] B.-N. Vo, B.-T. Vo, and M. Beard, "Multi-sensor multi-object tracking with the generalized labeled multi-bernoulli filter," *IEEE Trans. Signal Process.*, vol. 67, no. 23, pp. 5952–5967, 2019.

[47] S. E. Nordholm, H. H. Dam, C. C. Lai, and E. A. Lehmann, "Broadband beamforming and optimization," in *Academic Press Library in Signal Processing*. Elsevier, 2014, vol. 3, pp. 553–598.

[48] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128 837–128 868, 2019.

[49] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261–318, 2020.

[50] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[51] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "DSFD: dual shot face detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5060–5069.

[52] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 269–286.

[53] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (src) on a large-aperture microphone array," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 1. IEEE, 2007, pp. I–121.
[54] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.
[55] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pat. Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, 2000.
[56] B.-N. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo methods for multitarget filtering with random finite sets," *IEEE Trans. Aerospace and Electronic Systems*, vol. 41, no. 4, pp. 1224–1245, 2005.
[57] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 5. IEEE, 2001, pp. 3021–3024.
[58] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 826–836, 2003.
[59] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3291–3304, 2006.
[60] J. Benesty, C. Paleologu, T. Gänsler, and S. Ciochină, "Recursive least-squares algorithms," in *A perspective on stereophonic acoustic echo cancellation*. Springer, 2011, pp. 63–69.
[61] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3447–3457, 2008.
[62] M. Beard, B. T. Vo, and B.-N. Vo, "A solution for large-scale multi-object tracking," *IEEE Trans. on Signal Process.*, vol. 68, pp. 2754–2769, 2020.
[63] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoust., Speech, and Signal Process. Proc. (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.
[64] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.
[65] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2046–2057, 2011.
[66] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.